NYC Data

Schools Pipeline

Project Architecture & Data Lifecycle

PRESENTED BY

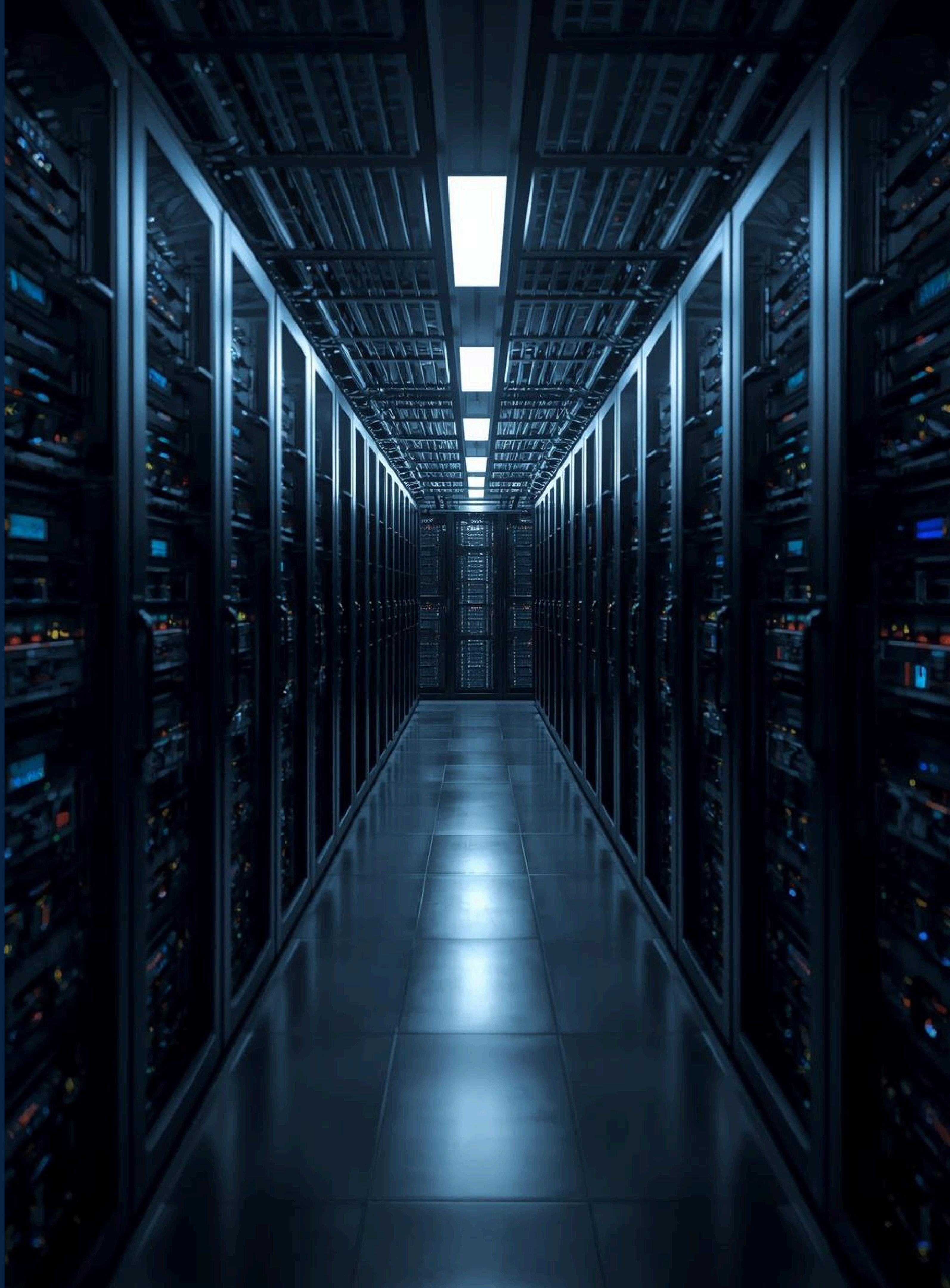
Anny H. LLosa

Data Ingestion

The **data ingestion** process for the NYC Schools Data Pipeline utilizes CSV datasets from NYC Open Data to ensure accurate and comprehensive analysis of school safety and directory information.

Storage and Relational Logic

This section reveals how **PostgreSQL** is utilized as a robust relational database, ensuring data integrity and consistency across various tables through structured relational logic.



ETL Processing: Automation and Data Cleaning

AUTOMATED FILE HANDLING

We utilize the **Python glob library** to create dynamic file handling mechanisms, ensuring efficient ingestion and processing of data files into the pipeline seamlessly.

DATA CLEANING PROCESS

Our ETL pipeline processes **28,151 safety incident records**, employing rigorous validation techniques to enhance data accuracy and reliability throughout the lifecycle of the project.

SAT RECORDS VALIDATION

We validate **478 SAT records** through automated checks, reinforcing the data quality and consistency necessary for meaningful insights and reporting in our NYC Schools Data Pipeline.

Contact Us

ANNY H. LLOSA

https://github.com/AnnyLlosa/nyc-schools-analysis

EMAIL ADDRESS

einny21@hotmail.com